Contrast Classes and Agreement in Climate Modeling

Corey Dethier

[forthcoming in *European Journal for Philosophy of Science*]

Abstract

In an influential paper, Wendy Parker argues that agreement across climate models isn't a reliable marker of confirmation in the context of cutting-edge climate science. In this paper, I argue that while Parker's conclusion is generally correct, there is an important class of exceptions. Broadly speaking, agreement is not a reliable marker of confirmation when the hypotheses under consideration are mutually consistent—when, e.g., we're concerned with overlapping ranges. Since many cutting-edge questions in climate modeling require making distinctions between mutually consistent hypotheses, agreement across models will be generally unreliable in this domain. In cases where we are only concerned with mutually exclusive hypotheses, by contrast, agreement across climate models is plausibly a reliable marker of confirmation.

0 Introduction

Climate scientists often use groups—called "ensembles"—of models to provide evidence for hypotheses about the past, present, and future climate. Sometimes these models all "agree" on a hypothesis; they all generate a result that indicates that said hypothesis is true. What should we conclude in such cases? In an influential paper, Wendy Parker (2011, 2018) argues that agreement across climate models does not provide significant confirmation, at least not where hypotheses about the future are concerned. Or, as she carefully puts the point: "When today's climate models agree that an interesting hypothesis about future climate change is true, it cannot

be inferred—via the arguments considered here anyway—that scientists' confidence in the hypothesis should be significantly increased" (Parker 2018, 275).

Parker's conclusion seems to be widely endorsed in the literature.[1] Complicating the picture is that much of literature employs concepts of "robustness" that are distinct from Parker's "agreement" and robustness has been interpreted in a wide variety of ways. So, for example, Lloyd (2015b, 60) explicitly endorses Parker's conclusion, but argues that hypotheses are confirmed when robust in her preferred sense—a point that O'Loughlin (2021) extends. Winsberg (2021, S5118) is similarly explicit in agreeing with Parker's conclusion narrowly construed but argues that the situation is different when we expand our focus beyond climate models to robustness over other sources of evidence.

In this paper, by contrast, I want to reconsider the more narrow construal of Parker's arguments; while I'm sympathetic to the positions of Lloyd (2015b), O'Loughlin (2021), and Winsberg (2021), I think there's more to be said about the narrow case. Ultimately, I'll argue that Parker is right: *generally speaking*, that an ensemble of climate models agrees on an "interesting" hypothesis does not provide good grounds for a significant increase in confidence. Importantly, however, there is an easily-identifiable subset of cases in which agreement is plausibly a marker of confirmation, and the general conclusion is more a function of the relationship between "agreement" and the kinds of questions that climate scientists deem "interesting" than it is a sign of a defect in the models.

To get a more precise feel for the argument, consider the following two questions concerning equilibrium climate sensitivity (ECS), or the amount that mean global temperatures will increase given a doubling of atmospheric CO_2 concentration relative to pre-industrial levels:

> Is ECS within the range 1.5-4.5°C?
>
> Which temperature range does ECS fall within?

The first of these two questions simply asks whether we should predict that a quantity of interest (ECS) will be inside of a given range or outside of it. In other words: the hypotheses are *mutually exclusive*. As we'll see, it's plausible that if every model in an ensemble agrees that ECS will be inside this range, that warrants an increase in confidence that the true value is in the range rather than outside of it—depending on one's priors, perhaps even a significant one. By contrast, the second question asks us to make a choice between different, potentially overlapping, ranges. Here the

[1] For a sampling of the discussion, see Baumberger, Knutti, and Hadorn (2017), Frigg, Thompson, and Werndl (2015), Gluck (2023), Harris (2021), Harris and Frigg (2023), Justus (2012), Lloyd (2015b), O'Loughlin (2021), and Winsberg (2021).

hypotheses are *mutually consistent*: the truth of one does not (necessarily) rule out the others. On extremely general grounds, it is *not* plausible that if every model in an ensemble agrees that ECS will be inside a particular range, that warrants an increase in confidence that the true value is in that range rather than in a different, overlapping, range.

As this example makes clear, the contrast class—that is, the way we divide up the hypothesis space—matters.[2] After demonstrating the importance of contrast classes, I'll show that cutting-edge climate science often involves contrasts where alternative hypotheses are mutually consistent. Together, these two facts vindicate Parker's general conclusion while also indicating that there is an important and easily identified class of cases where agreement should increase our confidence.

I begin with an abstract discussion of contrast classes, confirmation, and the nature of agreement (§1). The main body of the paper examines the application of these general principles to climate modeling. First, I argue that agreement is capable of distinguishing between mutually exclusive hypotheses under extremely weak conditions, and that we have little reason to think that these conditions fail in the climate modeling context (§2). Second, I show that there are formal reasons why agreement is ineffective at distinguishing between mutually consistent hypotheses (§3). Finally, I survey the "interesting" questions at the cutting-edge of climate modeling, arguing that while many require distinguishing between mutually consistent hypotheses, some don't—meaning that while Parker's conclusion is generally right, there is an important class of exceptions (§4).

1 Contrast classes, confirmation, and agreement

It's helpful to begin with the two questions from the introduction. The first of these two questions, namely,

> Is ECS within the range 1.5-4.5°C?

has two possible answers: (1) ECS *is* within the range 1.5-4.5°C and (2) ECS *is not* within the range 1.5-4.5°C. This question concerns only whether or not some proposition is true and so the potential answers are a proposition and its negation; these answers are *mutually exclusive*. By contrast, the second question,

> Which temperature range does ECS fall within?

[2]Lloyd (2015a) makes a similar point on very different grounds.

has infinite potential answers, including: ECS is within the range 1.5-4.5°C, ECS is within the range 3.0-6.0°C, ECS is within the range 1.0-5.0°C, and etc. Unlike the first question, this second question asks us to distinguish between overlapping, and thus *mutually consistent*, hypotheses. While one and the same proposition—namely, ECS is within the range 1.5-4.5°C—serves as a potential answer for both questions, the *contrast class*—that is, the set of other hypotheses that we're comparing this proposition to—differs in the two cases.

Generally speaking, answering either of these questions requires showing that one of the hypotheses is preferable to the other members of the contrast class. As a consequence, a given hypothesis may be preferable in one setting but not in others: ECS is within the range 1.5-4.5°C can be preferable to its negation without being preferable to the hypothesis that ECS is within the range 1.0-5.0°C. Since truth (or the probability thereof) is one desideratum for a preferable hypothesis, one way that evidence can help us answer a question is insofar as it confirms one of the potential answers in the contrastive sense of raising its probability relative to the alternatives.[3]

In climate science, as in many real life cases, we face a problem in determining whether and to what degree evidence supports one hypothesis rather than another. So, as Parker (2010) and others (e.g. Jebeile and Barberousse 2021) have emphasized, climate models both rely on risky assumptions and idealizations and are "opaque" in the sense that the implications of these assumptions and idealizations are (extremely) hard to evaluate. These facts make it hard to estimate how accurate any model is likely to be and thus to tell whether a particular model report confirms a particular hypothesis. On these grounds, both philosophers and climate scientists have argued against the application of standard statistical tools for evaluating the implications of evidence.[4] After all, the problem is essentially that we don't know the precise likelihood of a given model report on different hypotheses. Given that we lack this knowledge, the formal approaches that require it are unwarranted or inapplicable.

Heuristics—rules for preferring an answer that rely on markers that are more readily apparent than formal probabilities are—are one promising alternative to formal

[3]Why focus on probability-*raising* here? Because some of the other desiderata for a good hypothesis—such as informativeness or decision-relevance—trade off with posterior probability. For instance: the hypothesis that ECS is in the range 1.0-5.0°C will *always* be more probable but less informative than the hypothesis that it is in the range 1.5-4.5°C. What we want from our evidence is information about how much more probable—that is, we want it to raise the probability of one hypothesis relative to the other (see §3 for further discussion).

[4]There are a variety of different views in this vicinity, each with their own caveats and alternatives. See, e.g., Carrier and Lenhard (2019), Jebeile and Barberousse (2021), Katzav et al. (2021), Parker and Risbey (2015), Stainforth et al. (2007), and Winsberg (2018). For a contrasting view, see Dethier (2022, 2023a).

probabilistic reasoning. One potentially useful marker is *agreement*: given a number of different sources of evidence, all of them could "agree" that a hypothesis is true, and this agreement might give us reason to prefer the agreed-upon hypothesis. Following Parker, we can understand what agreement means in the climate modelling context as follows:

> When does an ensemble agree that a hypothesis is true? Assume that the values of model variables can be translated into statements regarding target system properties. Then a simulation indicates the truth (falsity) of some hypothesis h about a target system if its statements about the target system entail that h is true (false). For example, if h says that temperature will increase by between 1°C and 1.5°C, and each of the simulations in an ensemble indicates an increase between 1.2°C and 1.4°C, then each of those simulations indicates the truth of h and the ensemble is in agreement that h is true. (Parker 2018, 290–91, fn 2)

A gloss: on the view we're adopting from Parker, agreement is a matter of shared content. That is, whether or not there's "agreement" between models depends on what those models say, in much the same way that whether or not there's "agreement" between witnesses depends on what those witnesses say (though, of course, people and models don't quite "say" things in the same way). Presumably, it is easier to identify whether two models "say" that a hypothesis is true than it is to determine the likelihood of the evidence given various different hypotheses, meaning that agreement might potentially help us answer questions in cases—like that of climate modeling—where quantitative likelihoods are not available and qualitative heuristics must be employed instead.[5]

Whether agreement is actually a useful heuristic in climate science depends on whether it is a reliable marker of confirmation. That is, where there is agreement to be found, it must be the case that the agreed-upon hypothesis is (almost always) confirmed in the formal probabilistic sense. In our context, in other words, agreement is a reliable marker of confirmation iff in most cases where climate models agree on a particular h_1, the odds of h_1 relative to the members of its contrast class increase.[6]

[5]This treatment of agreement should be distinguished from those in which agreement (or a related concept like "robustness") is interpreted as a particular probabilistic condition (see, e.g., Dethier forthcoming; Schupbach 2018). Since I'm following Parker closely here, I'm adopting her approach, but in other contexts an alternative understanding of agreement may be preferable.

[6]We can model Parker's concept of agreement as follows. Let s_i represent what I'm calling the estimate and what Parker calls the "statement" of the i^{th} model. Then two models agree on h_1 iff $s_1 \vDash h_1$ and $s_2 \vDash h_1$. Note that the evidence (e_i) in this situation is not s_i but the fact that the

When is agreement a reliable marker of confirmation in this sense? Well, evidence e confirms hypothesis h_1 relative to h_2 iff the likelihood of the evidence on supposition of the first hypothesis ($Pr(e|h_1)$) is greater than the likelihood of the evidence on supposition of the second hypothesis ($Pr(e|h_2)$) (Chandler 2013). Following Myrvold (1996, 2017), notice that the likelihood ratio for any set of evidence $E = \{e_1, ..., e_n\}$ can be rewritten in terms the likelihood ratios of the individual pieces of evidence as follows:

$$\frac{Pr(E|h_1)}{Pr(E|h_2)} = \prod_{e \in E} \frac{Pr(E|h_1)}{Pr(e|h_1)} \times \prod_{e \in E} \frac{Pr(e|h_2)}{Pr(E|h_2)} \times \prod_{e \in E} \frac{Pr(e|h_1)}{Pr(e|h_2)}$$

For simplicity, call the first and second terms in this equation $1/V(E, h_1)$ and $V(E, h_2)$, respectively. Each of these V terms measures the degree to which E departs from probabilistic independence given the relevant hypothesis: the larger V is, the more "varied" the evidence is on the hypothesis. Essentially, this variation metric measures how surprising it is to observe the different pieces of evidence together. For instance, it might not be surprising to learn that a particular person has either a cat or a dog, but quite surprising to learn that they have both; it's this interactive effect that the V terms measure.

Together, the two V terms give the variation of the total evidence set given the hypothesis h_2 over the variation of the same evidence given h_1. The equation just given can then be rewritten as:

$$\frac{Pr(E|h_1)}{Pr(E|h_2)} = \frac{V(E, h_2)}{V(E, h_1)} \times \prod_{e \in E} \frac{Pr(e|h_1)}{Pr(e|h_2)} \tag{JL}$$

Myrvold glosses the insight contained in (JL) as follows: "a diverse body of evidence confirms a hypothesis more strongly if the hypothesis renders the evidence less diverse" than the negation of the hypothesis does (Myrvold 1996, 663). Dethier (forthcoming) notes that "we can add that in fact the relationship is really one of *to the degree that* rather than merely *if*"; if we hold fixed the individual likelihood ratios, the joint likelihood will vary directly with the ratio between the two V terms (see also Schlosshauer and Wheeler 2011; Wheeler and Scheines 2013).[7]

i^{th} model "says" that s_i. After all, within this idealized setting, we can assume that we know what the model says, but we can't update on the truth of that content because we don't know whether "what the model says" is true. Agreement is then a reliable marker of confirmation to the extent that agreement ($s_1 \vDash h_1$ and $s_2 \vDash h_1$) makes confirmation ($Pr(e_1 \& e_2|h_1) > Pr(e_1 \& e_2|h_2)$) likely.

[7] Of course, if we view the formalism here itself as a model of the testing situation, we might reasonably only draw qualitative conclusions about "real" degree of confirmation. My point here is that formal measures of confirmation track the ratio between the two V terms.

As this analysis reveals—and as the formal literature on the subject has repeatedly stressed—"independence" strictly understood has no *special* role in confirmation.[8] What matters is the degree of variation between the sources of evidence as measured by the ratio between V terms. Of course, we should expect that informal or intuitive markers of independence—such as whether or not two experiments employ the same instruments—will track this formal quantity. As we'll see, however, the connection between the intuitive markers and the formal conditions on confirmation is not straightforward, and confirmation is possible even in situations where there's a substantial intuitive connection between the different sources.

Effectively, (JL) provides us with an extremely general formal model of confirmation in the setting of multiple sources of evidence—the setting we find ourselves in when evaluating the evidence provided by groups of climate models. This model is useful in the present setting because it provides two precise conditions that are jointly sufficient for the confirmation of h_1. If we can show that agreement is highly correlated with these conditions, we will have established that agreement is a reliable marker of confirmation.

The two conditions are as follows. First, say that a set of evidence E "converges" on h_1 (relative to h_2) iff every piece of evidence raises the probability of h_1 relative to h_2. Second, say that a set of evidence is "weakly varied" (with respect to h_1, h_2) iff the ratio between $V(E|h_1)$ and $V(E|h_2)$ is not very large. Recall our intuitive gloss of the V terms; they measure how surprising it is to observe all the evidence together. So what we don't want is that the evidence is *substantially* more surprising given h_1 than given h_2. What "substantially" means here will depend on context. More precisely, we'll say that the evidence is "weakly varied" iff the ratio between $V(E|h_1)$ and $V(E|h_2)$ is not larger than the product of individual likelihoods. Notice: give convergence, the product of the individual likelihoods will be larger than 1, often much larger. This means that the models can be highly correlated—read: observing the reports together is very unsurprising—and still be weakly varied; so long as h_2 isn't a *much* better explanation of their correlation than h_1, we'll still see confirmation.

Our two conditions are then:

E *converges* on h_1 iff for all $e \in E$, $Pr(e|h_1) > Pr(e|h_2)$.

E is *weakly varied* iff $V(E|h_1)/V(E|h_2) < \prod_{e \in E}[Pr(e|h_1)/Pr(e|h_2)]$.

[8] Dethier (forthcoming), Myrvold (1996), and Wheeler (2012) all make this point explicitly using variations on the same formalism adopted here. The point holds in other settings, however, as illustrated by the work of Bovens and Hartmann (2003) and Schupbach (2018).

As just noted, these two conditions are jointly sufficient for confirmation: if E converges on h_1 relative to h_2 and is weakly varied with respect to h_1 and h_2, then E confirms h_1 relative to h_2. These conditions are, notably, fully general: they hold regardless of the relationship between h_1 and h_2, the nature or source of the evidence that makes up E, etc.

The question we have to ask, then, is whether agreement across climate models reliably correlates with these two conditions. In the next two sections, I'll argue that the answer to this question is that it depends on the contrast class: when we're concerned only with mutually exclusive hypotheses, it's plausible that agreement does reliably correlate with the two conditions that are sufficient for confirmation; when we're concerned with mutually consistent hypotheses, by contrast, it isn't.

2 Agreement and mutually exclusive hypotheses

To determine whether our two formal conditions correlate with agreement, we need to first map these conditions onto the practice. So let h_1 be the hypothesis that ECS is within the range 1.5-4.5°C and, because we're currently interested in the setting where the two hypotheses are mutually exclusive, the contrasting hypothesis is $\neg h_1$, or ECS is not within the range 1.5-4.5°C. The pieces of evidence that we're updating on, e_1, e_2, e_3 etc., are facts about the outputs of the models: that the first model output an estimate for ECS of x, that the second model output an estimate for ECS of y, etc. Notice, importantly: this means that the relationship between the first and second model factors into the probabilistic relationship between e_1 and e_2; if the two models are identical, then $Pr(e_2|e_1) = 1$. In other words, what our hypotheses must explain is the fact that each model yields the output that it does. Roughly speaking, the question is whether "agreeement" between the models—i.e., the outputs all falling within a particular range—is better explained by the true value falling within that range (and the models tracking it) or by some sort of shared error.[9]

In this restricted setting where we're only concerned with mutually exclusive hypotheses, both convergence and weak variation are extremely minimal conditions, and if each of the models in an ensemble agrees on one of the two hypotheses in

[9]Of course, putting the question in these terms is potentially misleading in one respect: the ranges for ECS are not put forward ahead of time as hypotheses; instead, they're inferred—typically using statistics—from the outputs of models and other evidence. My discussion here should be understood as representing the formal support relationships between the evidence and the space of hypotheses that we could draw inferences about on the basis of that evidence and not as capturing anything like the temporal order involved in the practice of drawing these inferences.

the sense of generating an estimate for ECS that falls within its range, then we should *expect* that these conditions are met—or, more minimally, the arguments in the literature are not sufficient to motivate rejecting this expectation. And thus, we don't have good reason to reject the claim that agreement on h_1 is a reliable indicator of the confirmation of h_1 relative to $\neg h_1$.

Let's begin with convergence, which is the more intuitive condition. Essentially, what is required for agreement to be a reliable marker of convergence is that if all the models "say" that a hypothesis is true in the sense of delivering an estimate that falls within the range picked out by that hypothesis, then the evidence that is generated by each model is more probable on the assumption that the true value falls within this range than on the assumption that it doesn't. So in our running example, each of the models must be more likely to deliver an estimate for ECS that falls within a given range (e.g., 1.5-4.5°C) when the truth also falls within that range then when the truth falls outside of it. Accordingly, if the condition fails to hold, that means that there is at least one model that is so seriously inaccurate as to be *anti-correlated* with the truth: when we learn that it delivers an estimate that falls between 1.5 and 4.5°C, we should *lower* our confidence that true value for ECS is actually in this range.

Notice how strong this requirement is. The demand here is not just that there is one model or another that is very untrustworthy in the sense that we should assign very little confidence to its estimates. It's that one of the models is untrustworthy in the sense that we expect it to be less reliable than flipping a coin. It's highly implausible that any given model is this unreliable. It's even more implausible that a given model would be this unreliable without there being substantial evidence of its unreliability—after all, these models are thoroughly studied and vetted as part of the construction process, in intercomparison projects (O'Loughlin 2023), and in subsequent studies that compare their results to present-day climate data (see, e.g., Eyring et al. 2020). Absent evidence that a specific climate model is worse than chance with respect to a particular variable, therefore, we should presume that the climate models that make up standard ensembles converge in the very minimal sense offered here.

In her discussion of agreement in climate modeling, Parker offers specific arguments that she takes to undermine a condition that is equivalent to convergence (Parker 2018, 282, 288). To summarize: today's ensembles suffer both from the omission of potential relevant processes and from (shared) idealizations about the nature of processes that are risky or unjustified. Further, attempts to evaluate how these omissions and idealizations affect the predictions of ensembles have indicated that there's a correlation between one model giving an (in)accurate prediction / estimate

and the next one doing the same.[10]

I have no grounds for disputing Parker's descriptive claims about the defects of contemporary models. What I dispute is that these defects give us reason to reject convergence on h_1 relative to $\neg h_1$ in cases where all of the models agree on h_1. To see why, consider the two main problems that Parker raises for contemporary climate models, namely that (a) they (all) omit some processes and (b) they idealize others (often all in the same way). These are good reasons for thinking that the models are imperfect, that they are likely not to give answers that are accurate to the level of precision that we might desire. They are also good reasons for thinking that simply collecting more models that share these same defects will not provide definitive answers to our questions, because there will be potential sources of error shared across all of the models.

Nevertheless, neither the omission of some processes nor the idealization of others is a good reason for thinking that a model or ensemble is so defective that they are more likely to provide evidence for false hypotheses than for true ones. Generally speaking, if an idealized model estimates that some quantity is x, we think that this result makes it (more) likely that the true value falls in $x \pm y$, where y is the relevant margin of error. In good cases, y is very small; in bad cases, it's very large. Idealizations and omissions are generally speaking a reason to increase y or to shift it in one direction or another (i.e., to think that the true value is between $x - u$ and $x + v$), but not a reason to think that the model's estimate of x makes it more likely that the true value is radically different from x. The same is true of shared omissions and idealizations: the presence of shared idealizations across models does not generally make it more likely than not that those models are all wrong.[11]

Of course, that's not to say that there aren't special cases where idealizations or omissions could undermine our confidence in convergence. So, for example, imagine that we knew that omitting some process had a massive *biasing* effect: the omitted process generates a net positive feedback, and including that process in the model would increase the estimate of ECS by 3°C.[12] Then an estimate of 2°C would give us good reason to think that the true value of ECS is outside the range 1.5-4.5°C. And so we would have a good reason to doubt that agreement on this range is a

[10]Though that doesn't mean that they're inaccurate in the same way. On the contrary, there's a known correlation between error and spread: when the models get a prediction wrong, they tend to get it wrong in a bunch of different ways (see Knutti et al. 2010).

[11]My point here is not that we should add error bars around h_1 in this case, but that the mere presence of idealizations doesn't undermine the convergence condition except in special cases; see the next paragraph.

[12]There are cases like this in regional climate modeling; see, e.g., Boé et al. (2020) and Schwingshackl et al. (2019).

good marker of convergence. But the point made above still holds: absent specific evidence to the contrary, we don't have good reason to expect that these kinds of extreme errors are present in the model. This point is particularly true in contexts where the models have been thoroughly studied and validated against available data. Convergence fails only when there's something seriously wrong with the models, and while it's always possible that validation studies could have missed an error, that possibility is not sufficient to motivate the conclusion that the models are so inaccurate that convergence fails.

Similar comments apply with regard to weak variation. Recall that we can think of the variation measure V as quantifying how surprising it is to observe the different pieces of evidence together. What we then require is that when all the evidence falls within the range specified by h_1, h_1 doesn't render this confluence of evidence *substantially* more surprising than $\neg h_1$ does. If observing the different pieces of evidence together is even roughly as surprising given both hypotheses, weak variation is satisfied. The same point about the implausibility of the models being worse than chance holds here as well, just at one level removed. While a failure of weak variation doesn't amount to the inclusion of a single model that is worse than chance, it does amount to the use of a group of models that are *jointly* defective in the sense that they're much more likely to agree on falsehoods than on truths. And there's no reason to think that extant ensembles are defective in this way.

The presence of omissions and idealizations—even quite serious ones—in the ensemble certainly does not provide such a reason. These sorts of flawed assumptions might make it so the evidence is equally surprising on both h_1 and $\neg h_1$; if the agreement between models is to be attributed to the failure of a particular idealization, then it doesn't matter what the true value of ECS is. But the presence of omissions and idealizations isn't a reason to think that the evidence is (substantially) less surprising on $\neg h_1$ than on h_1.[13] For that we need reasons to think that there are specific, confounding, interactions between the different pieces of evidence, interactions like those found in Stegenga and Menon (2017) or cases of Simpson's paradox. It's hard to image how such interactions would arise in the climate modeling context, however, let alone how they would arise without showing up in validation studies: if the models are really so constructed that they're substantially more likely to agree on

[13]As one reviewer stressed, one can imagine extreme scenarios—where, e.g., h_1 specifies ECS to five significant digits—where we should expect some idealization explains the agreement. I agree, but think the explanation for our intuition here is that the prior for the relevant h_1 is extremely low and the observed agreement is extremely surprising whether or not h_1 is true—it's not that $\neg h_1$ offers any better explanation of it that h_1 does. Indeed, our confidence in h_1 in this scenario should increase! Just not enough to overcome the low prior.

falsehoods than on truths, we would expect a correlation between agreement and error. In fact, what we see is the opposite: error is highly correlated with model spread (Knutti et al. 2010).[14] Just as is true of convergence, in other words, we have no reason to expect that the models actually employed in climate science are generally this defective, and good reason to think that if they were, we would know about it.

Perhaps the foregoing arguments are unfair, however, given that the target is Parker's view. Rather than arguing that we should think that convergence or weak variation fails, an opponent might argue that, given the criticisms raised by Parker, we should suspend judgment on the status of the connection between agreement and these conditions and thus also on whether agreement is reliable marker of confirmation. In fact, Parker indicates that this is her view with respect to convergence: "The claim here is not that individual modeling results have negative evidential relevance but that their evidential status (with regard to interesting hypotheses about long-term climate change) is largely unknown" (Parker 2018, 293, fn 28). I'm tempted to respond that we should suspend judgment here if and only if we are *roughly* equally confident as not that agreement positively correlates with the conditions given above. What I've essentially been arguing is that we have good reasons for expecting correlations that are at least weakly positive.

Tempting as that line is, however, the Parker-inspired critic would be correct in pointing out that it's too fast. We can agree that an agent ought to suspend judgment only when the probability of convergence given agreement is "roughly" .5 while disagreeing on what exactly "roughly" means here. That is, we can imagine two agents who disagree about whether or not to suspend judgment because they differ in how cautious they want to be in their positive epistemic judgments. The arguments I've given motivate the following: given the actual practices involved in model construction and validation, we should expect that the models of climate science are *not* anti-correlated with the truth. Minimally, therefore, a reasonable agent (a) *cannot* conclude that agreement doesn't provide evidence and (b) *need not* suspend judgment. The arguments given in the literature don't justify the former and don't force the latter. At worst, the judgment that agreement provides evidence is defensible and a reasonable agent might adopt it without being seriously incautious.

This brings us to another respect in which the arguments just given might be seen as unfair: Parker, at least, is concerned with "significant" increases in confidence, while all that I've established is that it is reasonable to increase one's confidence to

[14]The claim here is: where climate scientists haven't found these interactions, we don't have sufficient reason to expect them to exist. The restriction is important. In other contexts, we might expect that such interactions are common enough that even experts miss them with a high degree of frequency. I see no reason to postulate that that's the case here, however.

some degree.[15] Of course, the point of the last paragraph applies here too: there's no bright line for what counts as significant. Nevertheless, it's worth stressing how easy it is for agreement to generate evidence that is significant in an intuitive sense. Consider a toy example characterized by the following three conditions. First, our ensemble consists of 10 models (around the average size of extant ensembles). Second, each of these models is about 1.5 times as likely to say that h_1 is true if it is than if it isn't—that is, the evidence provided by an individual model favors h_1 but is "not worth more than a mere mention" (Kass and Rafferty 1995, 777). Third, the convergence of models is five times as surprising given h_1 as it is given $\neg h_1$; before carrying out our simulations, we're quite a bit more skeptical that the models will agree for good reasons then for bad.

Recall that (JL) tell us that the joint likelihood ratio is equal to the *inverse* of the ratio between V terms (.2) times the product of the individual likelihood ratios (1.5^{10}), meaning that joint likelihood ratio is around 12 ("strong" evidence according to Kass and Rafferty 1995, 777). A standard Bayesian agent who was indifferent between h_1 and $\neg h_1$ prior to encountering the ensemble should now have a relative confidence in h_1 over .9, which is surely a significant increase in confidence. This example is (highly) arbitrary, but the assumptions are overly pessimistic if anything: we've assumed a realistic number of models, each of which provides poor evidence, and that are on the whole much worse than a similar ensemble of "independent" models. And yet the result is significant confirmation. Agreement is an extraordinarily powerful tool for deciding between mutually exclusive alternatives.

Allow me to step back. What I've argued in this section is that agreement across even highly flawed ensembles can be a reliable marker of confirmation when we're concerned only with mutually exclusive hypotheses. As is hopefully already clear, I don't take my arguments to definitively show that we should significantly increase our confidence in all such cases—there are times when we know that a particular set of models is biased or misleading. Where multiple sources of evidence agree on one of a set of mutually exclusive hypotheses, however, the conditions sufficient for confirmation—even significant confirmation—are *so* weak that we don't have good reason to reject them absent this kind specific evidence about the individual models. At minimum, more arguments are needed to show that these conditions should typically be rejected, but I think the stronger conclusion is actually warranted: typically, agreement among models on one of a set of mutually exclusive hypotheses is powerful evidence for the truth of that hypothesis. The reason why is illustrated by the arbitrary but pessimistic toy example discussed above: a set of different model

[15] Reasonably, Parker doesn't define "significant," but she does note that it depends on contextual factors (Parker 2018, 281).

reports can be very informative evidence even when the individual reports are both poor evidence by themselves and very interrelated. The same is not true in the context of mutually consistent hypotheses, as I argue in the next section.

3 Agreement and mutually consistent hypotheses

In this section, I argue on general grounds that agreement is not a reliable marker of confirmation in the context of mutually consistent hypotheses.

Recall: our goal is to show that one of a set of mutually consistent hypotheses—hypotheses that might place ECS in the range of 1.0-5.0°C, 2.0-5.0°C, 4.0-6.0°C, etc.—is to be preferred over the others; since we're focusing only on the desire for true hypotheses and not other virtues that a hypothesis might have, agreement on a hypothesis indicates that the hypothesis preferable iff agreement is a reliable marker of an increase in the odds of that hypothesis relative to the other hypotheses under consideration.

Agreement is *not* a reliable marker of confirmation in this context, however. Intuitively, the problem is quite simple. Consider two overlapping potential ranges for ECS, such as 1.5-4.5°C and 1.0-5.0°C. Suppose that a given ensemble agrees on the first range in the sense that every model generates an estimate that falls within that range. Then every model generates an estimate that falls within the latter range as well. So the ensemble agrees on both ranges. But the two ranges cannot both be confirmed relative to the other: only one of the two agreed-upon hypotheses can see its relative odds increase. The point this case illustrates is that agreement cannot be a reliable marker of relative confirmation when there's agreement on both hypotheses. The problem is that when working with contrast classes that include mutually consistent hypotheses, the models are liable to agree on multiple hypotheses in the requisite sense. Since these hypotheses cannot all be confirmed relative to each other, the upshot is that agreement is not a reliable marker of confirmation.[16]

[16]Of course, there's no guarantee that the models will agree on all of the mutually consistent hypotheses that we're interested in. In that case, by the reasoning of the last section, there will be some confirmation of the set of hypotheses that the models do agree on. But notice: if the models all fall within the range 1.0-5.0°C, my arguments in the last section *strictly* only warrant increased confidence that ECS falls within that range relative to the hypothesis that it doesn't. Plausibly, my arguments could be extended to confirmation relative to a fully excluded range, like 6.0-10.0°C. But they don't extend to confirmation relative to 1.0-10.0°C. As we'll see in the next section, these kinds of overlapping hypotheses with what are effectively different "error bars" are one of the major cases that Parker is concerned with. Or, in other words, in the cases at issue, our prior probability is often sufficiently concentrated that at least some results fall within the scope of each of the interesting hypotheses. Where that isn't true, however, I'm happy to acknowledge that we have confirmation

Notice that the difficulty here arises from the fact that agreement comes apart from confirmation in a way that makes agreement less apt for deciding between mutually consistent hypotheses. So consider again the two ranges 1.0-5.0°C and 1.5-4.5°C. Just as evidence that agrees on the latter agrees on the former, so too is the former guaranteed to have at least as high a posterior probability as the latter. With agreement, however, this is all that there is to say about the issue; because agreement is binary, it marks no difference between the two hypotheses. Confirmation, by contrast, can mark a difference: the odds ratio between these two ranges is not fixed, meaning that evidence can confirm the narrower relative to the wider. So while confirmation theory cannot tell us that the narrower hypothesis is true but the wider one false (nothing can do that, given that the narrower entails the wider), it can tell us that the narrower is a better representation what the evidence supports (compare Chandler 2007). Agreement has no resources to do the same.

To illustrate the problem, it's helpful to return to the formal model from note 6. In that context, the problem is that our notion of agreement allows for weakening the "consequent" and confirmation doesn't.[17] So two estimates s_1 and s_2 agree on a hypothesis h_i just in case the truth of any one of the estimates entails h_i—e.g., $s_1 \vDash h_i$ and $s_2 \vDash h_i$. Trivially, however, if h_i is entailed, then so is $h_i \vee h_j$ for any hypothesis h_j. Indeed, if we write the disjunction of all the estimates out as $s_1 \vee ... \vee s_n$, then the ensemble agrees on *any* hypothesis that can be rewritten as $s_1 \vee ... \vee s_n \vee X$. Confirmation doesn't work like this: that e confirms h_i does not guarantee that it confirms $h_i \vee h_j$—let alone that it does so *relative* to h_i. When our hypothesis space includes only one hypothesis that can be re-written as $s_1 \vee ... \vee s_n \vee X$, it's plausible that this hypothesis—the sole hypothesis that is consistent with any one of the statements of the ensemble being true—is confirmed. When the potential answers include more than one hypothesis that can be re-written in this way, by contrast, agreement cannot be expected to track confirmation.

The point just made explains why agreement is not generally useful in the context of mutually consistent hypotheses; it also helps explain why there seem to be notable exceptions to this general conclusion, such as Perrin's use of multiple measurements to estimate Avogadro's number (see Perrin 1916). In these cases, scientists have some other grounds for restricting the set of potential hypotheses so that there is only one answer equivalent to $s_1 \vee ... \vee s_n \vee X$. The most straightforward (and I suspect the most common) such grounds are well-founded expectations about experimental error.

---

of some subset of the mutually consistent hypotheses.

[17] The use of "consequent" here can be made more appropriate by reframing the discussion in terms of an agreement conditional and a confirmation conditional á la Joyce (1999), but I think that's unnecessary for communicating the point.

For instance, if we have reason to believe that every one of our estimates is accurate to within 1°C but no reason to think that any one of them is more accurate than that, then, from all of the infinity of ranges that the estimates agree on, we have grounds for preferring one of these, because there is at most one such that it includes all and only the values that are within 1°C of every estimate. Of course, in most cases, our expectations about experimental error will not be this precise, but the same general lesson applies: knowledge about the potential for experimental error gives us a reason to prefer one hypothesis to all of the others that the set of results agrees on.

This use of background knowledge is essentially what vindicates appeals to agreement like those found in Perrin's estimation of Avogadro's number.[18] Given what Perrin knew about the experiments in question, he was able to put relatively accurate and highly precise (for 1908) limits on what possible values for Avogadro's number each of the experiments allowed. With these limits in hand, the problem Perrin faced was less "which numerical range does Avogadro's number fall within" than it was "which (if any) is the *unique* range such that all and only the values within it are within the expected experimental error of every result?" That there is such a unique range—that the results agree in this sense of all being within (relatively) well-defined experimental error bounds of each other—is often extremely powerful evidence; it's a good indication that we should prefer that (unique) range, at least provisionally. But this reasoning only works because extensive background knowledge is employed to restrict (in our earlier language) the class of hypotheses so that only one of them could be rewritten as $s_1 \vee ... \vee s_n \vee X$. In other words, the logic of reasoning from agreement *within experimental error* is very different from the logic of reasoning from agreement *simpliciter*—the modifier is doing a lot of epistemic work.

Unfortunately, it isn't really possible to consistently approach agreement across climate models in the same way.[19] The reason why is that we don't typically have the requisite background knowledge of how the individual models work and thus of how accurate they are likely to be. The standard case in climate modeling is not analogous to Perrin's, where we have a deep understanding of the specific instruments and techniques involved and the resulting constraints on how much "random" error we can expect in a given estimate. We can't, that is, give a well-founded estimate of a value of y such that a given model is sufficiently likely to deliver estimates within y

[18]The following discussion is simplified, but a close evaluation of Perrin's work vindicates the general picture while complicating our understanding of how well Perrin himself succeed at the task (see Smith and Seth 2020, chapter 6).

[19]Though, as O'Loughlin (2021) and Winsberg (2018) stress, there are specific cases in which we can further constrain model outputs using knowledge from other sources of evidence.

of the truth—or at least we can't for any values of y that are small enough to allow us to answer interesting questions in this way. The difficulties involved in understanding the inner workings of the models has been well covered by others (see, e.g., Jebeile and Barberousse 2021; O'Loughlin 2023), and so I won't belabor the point here.

It's tempting to respond to this result by concluding that something has clearly gone wrong. After all, it's ridiculous to think that when a group of models agree that ECS is between 1.5 and 4.5°C, that gives us *no* reason to prefer the hypothesis that ECS is actually in that range to the hypothesis that ECS is between -100 and 100°C. I agree: I too have the intuition that *of course* the clustering of models within the smaller range provides us with a reason to prefer it to the latter, and I think that intuition is right. The takeaway of the present section is that the agreement heuristic is the wrong way to accommodate it: what gives us reason to prefer the smaller range is not "agreement" in an informal sense—that can't distinguish between the two hypotheses—but the distribution or variation found in the ensemble. Similar comments apply to other cases. Suppose we're considering mutually consistent hypotheses like ECS is between 1.5 and 4.5°C and ECS is between 3.5 and 10.0°C. Again, the actual distribution of model results matters; but if all of the models actually fall in both ranges, agreement isn't the right way to cash that out.[20]

4 Interesting questions in climate modeling

The last section argued that agreement is not a reliable marker of confirmation in distinguishing between mutually consistent hypotheses. Insofar as that's right, Parker's thesis will be *generally* correct so long as many of "interesting" questions about the future climate require us to distinguish between mutually consistent hypotheses. In this final section, I'll argue that this is the case, but that there are crucial exceptions where we should expect agreement to be more informative.

To begin, note that Parker means "interesting" as something of a technical term:

> By an interesting predictive hypothesis, I mean a hypothesis about the future that scientists (i) do not already consider very likely to be true or

[20]Why not? Agreement ignores question of distribution: if we're asking whether the set of estimates agrees on h, we're putting aside variation in how much the different estimates support h or where the different estimates fall within the range captured by h. In Bayesian terms, we're not updating on the total evidence. One way of viewing cases in which agreement is a good heuristic is that they're cases in which the distribution makes little difference and so updating only on agreement approximates the results that we would get if we updated on the total evidence. As one reviewer stressed to me, at face value this point supports the argument by Lloyd (2015b) and others that we should adopt a different understanding of "robustness."

> very likely to be false and (ii) consider a priority for further investigation. In climate science today, these are typically, but not always, quantitative hypotheses about changes in global or regional climate on the timescale of several decades to centuries. (Parker 2018, 290, fn. 1)

That is, what Parker means by "interesting hypothesis about future climate change" are those hypotheses that are interesting to *scientists*—the hypotheses about the future that are investigated at the cutting-edge of science.

To determine the implications of the above arguments for Parker's claims about interesting hypotheses, we need to adapt her two conditions to account for contrast classes. The best way to do this is to shift from talk of interesting *hypotheses* to talk of interesting *questions*: as we saw in section 1, different questions admit different sets of potential answers. So, say that a question is interesting in this cutting-edge sense iff scientists (i) do not already consider one of the answers much more likely to be true than the other(s) and (ii) consider determining which of the possible answers is true to be a priority for further investigation. Note that these conditions yield a relatively straightforward way of determining what questions count as interesting in climate modeling: we should examine (i) what questions the climate scientists regard as not yet having been definitively answered and (ii) we should examine what questions they are working to answer.

On the first count, consider what the IPCC says about our running example (equilibrium climate sensitivity) in the 2013 assessment report:

> Equilibrium climate sensitivity is likely in the range 1.5°C to 4.5°C (high confidence), extremely unlikely less than 1°C (high confidence), and very unlikely greater than 6°C (medium confidence). The lower temperature limit of the assessed likely range is thus less than the 2°C in the AR4, but the upper limit is the same. (IPCC Working Group 1 2013, 16)

Here we are confronted with the IPCC's standard two-tier system of expressing confidence in hypotheses. How exactly we should interpret these nested probability judgments is an interesting question in its own right, but it seems clear that, at the very least, the IPCC considers a claim like "Equilibrium climate sensitivity is likely in the range 1.5°C to 4.5°C" to be a hypothesis—a proposition to be tested, accepted and rejected, etc.[21] Insofar as the IPCC's answers are indicative of their questions, they seem to take questions like the following to be interesting:

[21] I'm aware of two extended treatments of these sorts of probability claims: Dethier (2023b) and Winsberg (2018). Both explicitly build this point into their account.

> Which temperature range is ECS likely to fall within?
>
> Which temperature range is ECS unlikely to fall within, but not "very" or "extremely" unlikely to fall within?

These questions are straightforwardly questions that admit mutually consistent hypotheses as potential answers: they're asking *which* range we should prefer rather than *whether-or-not* we should prefer a given range to its negation.[22] Realistically, they're also asking these questions within a restricted setting rather than in the unrestricted formulation given here: ECS is a well-studied quantity and so climate scientists can reasonably be seen as limiting their attention to ranges of values that are compatible with the prior research on the subject—a range 15-20°C (for instance) is not one of the live hypotheses for the IPCC, whereas ranges like 1.5-4.5°C, 1.0-4.5°C, 2.0-6.0°C, etc. are.

The ECS-related questions that climate scientists prioritize answering are similar. A survey of recent papers on the subject, for example, reveals that climate scientists prioritize at least the following kinds of research relating to ECS: research into estimating ECS itself, particularly that of combining estimates from models with other sources of evidence, as in Sherwood et al. (2020); research into methods for more precisely estimating ECS using climate models (Dai et al. 2020); research into the accuracy of model-generated estimates of ECS (Gregory et al. 2020); and research into how various processes—especially those about which there's substantial uncertainty—affect ECS (Dong et al. 2020).

Not all of these are hypotheses that are *directly* about future climate change, but all of them at least have consequences for our understanding of the future climate, and none of them lends itself to mutually exclusive hypotheses in a straightforward way. The first kind of research is essentially research into our running example: which range does ECS fall within? The others concern how precise and/or accurate we can expect our estimates to be—where we should make the cutoffs between "likely," "unlikely," and "very unlikely" scenarios. None of these categories involves the kind of yes-no distinction between mutually exclusive hypotheses that agreement can help us make.

But, of course, there are other examples where climate scientists are concerned with mutually exclusive hypotheses. In the most recent assessment report, for example, the IPCC says that "Annual global land precipitation will increase over the 21st century as [mean global temperature] increases (*high confidence*)" (IPCC

[22] Of course, the real question here is "*what* is the value of ECS?" which is plausibly a question where the answers are mutually exclusive. But the models don't agree on that question—they deliver a range of different answers—and so the agreement heuristic cannot help us there.

2021, 556). The question here is clearly of the yes-no variety, though—the yes-no question answered—they immediately move on to estimating ranges.[23] We can find similar questions asked throughout the report. For instance: relatively little is known about how climate change will affect monsoons, and one question that the IPCC raises—but does not answer—is whether "wet get wetter, dry get drier" is a good characterization of future monsoon seasons (IPCC 2021, 584). Plausibly, these are cases where model agreement should significantly increase our confidence—at least absent specific reasons to think that the models are unreliable with respect to these variables.

The upshot of this section is that many, but certainly not all, of the interesting questions in climate science involved distinguishing between mutually consistent hypotheses, and thus we cannot *generally* assume that agreement is a reliable marker of confirmation. As just noted, however, that doesn't mean that agreement is not a reliable marker of confirmation in some cases. On the contrary, we've seen in the foregoing that there is an easily-identifiable subset of cases in which agreement is plausibly a marker of confirmation.

One final note. Parker's sense of "interesting" doesn't (and isn't intended to) exhaust all of the ways that a question might be interesting in a more intuitive sense. So, for instance, a question—even one that scientists deem settled—might be interesting in virtue of being relevant to an important political or financial decision. Along these lines, the public is or was interested in yes-no questions like "Is humanity's contribution to climate change positive rather than negative?" even though this question is considered to be conclusively resolved by climate scientists and the "interesting" question is *how* positive humanity's contribution is. At face value, therefore, agreement across models can be a useful tool for science communication even in settings where it wouldn't be particularly useful in answering cutting-edge questions.

5 Conclusion

Recall where we started: Parker argues that agreement across climate models doesn't warrant a significant increase in confidence in the agreed-upon hypothesis. Various responses have granted the narrow claim while arguing that there are other senses of "robustness" that can be used to confer confirmation. In this paper, by contrast, I've addressed Parker's argument within the narrow scope of her notion of agreement, arguing that while Parker is right in the general case, there are important—and

[23]Notably the IPCC *doesn't* appeal to model agreement in answering this question; by all appearances, their reasoning is purely statistical.

easily-identifiable—exceptions: the reason why agreement isn't a reliable marker of confirmation has more to do with the heuristic itself and the kinds of questions that climate scientists are interested in then it does with features of the models.

Acknowledgements

My thanks to Wendy Parker, who helpfully provided thorough comments on earlier drafts of this paper on two separate occasions; Morgan Thomson; and a number of anonymous referees for comments on earlier versions of this paper.

Funding

Funding for this paper was provided in part by the Deutsche Forschungsgemeinschaft Project No. 254954344/GRK2073 and in part by the National Science Foundation under Grant No. 2042366.

References

Baumberger, Christoph, Reto Knutti, and Gertrude Hirsch Hadorn (2017). Building Confidence in Climate Model Projections: An Analysis of Inferences From Fit. *Wiley Interdisciplinary Reviews: Climate Change* 8.3: e454.

Boé, Julien et al. (2020). Large Discrepancies in Summer Climate Change over Europe as Projected by Global and Regional Climate Models: Causes and Consequences. *Climate Dynamics* 54: 2981–3002.

Bovens, Luc and Stephan Hartmann (2003). *Bayesian Epistemology*. Oxford: Oxford University Press.

Carrier, Martin and Johannes Lenhard (2019). Climate Models: How to Assess Their Reliability. *International Studies in the Philosophy of Science* 32.2: 81–100.

Chandler, Jake (2007). Solving the Tacking Problem with Contrast Classes. *British Journal for the Philosophy of Science* 58.3: 489–502.

— (2013). Contrastive Confirmation: Some Competing Accounts. *Syntese* 190.1: 129–38.

Dai, Aiguo et al. (2020). Improved Methods for Estimating Equilibrium Climate Sensitivity from Transient Warming Simulations. *Climate Dynamics* 54.11-12: 4515–43.

Dethier, Corey (2022). When is an Ensemble Like a Sample? 'Model-Based' Inferences in Climate Modeling. *Synthese* 200.52: 1–20.

— (2023a). Against "Possibilist" Interpretations of Climate Models. *Philosophy of Science* 90.5: 1417–26.
— (2023b). Interpreting the Probabilistic Language in IPCC Reports. *Ergo* 10.8: 203–25.
Dethier, Corey (forthcoming). The Unity of Robustness: Why Agreement Across Model Reports is Just as Valuable as Agreement Among Experiments. *Erkenntnis*.
Dong, Yue et al. (2020). Intermodel Spread in the Pattern Effect and Its Contribution to Climate Sensitivity in CMIP5 and CMIP6 Models. *Journal of Climate* 33.18: 7755–75.
Eyring, Veronika et al. (2020). Earth System Model Evaluation Tool (ESMValTool) v2.0 – an extended set of large-scale diagnostics for quasi-operational and comprehensive evaluation of Earth system models in CMIP. *Geoscientific Model Development* 13.7: 3383–438.
Frigg, Roman, Erica Thompson, and Charlotte Werndl (2015). Philosophy of Climate Science Part II: Modeling Climate Change. *Philosophy Compass* 10.12: 965–77.
Gluck, Steward (2023). Robustness of Climate Models. *Philosophy of Science* 90.5: 1407–16.
Gregory, Johnathan M. et al. (2020). How Accurately can the Climate Sensitivity to CO_2 be Estimated from Historical Climate Change? *Climate Dynamics* 54.1-2: 129–57.
Harris, Margherita (2021). The Epistemic Value of Independent Lies: False Analogies and Equivocations. *Synthese* 199: 14577–97.
Harris, Margherita and Roman Frigg (2023). Climate Models and Robustness Analysis – Part II: The Justificatory Challenge. In: *Handbook of the Philosophy of Climate Change*. Ed. by Gianfranco Pellegrino and Marcello Di Paola. Cham: Springer: 1–22.
IPCC (2021). *Climate Change 2021: The Physical Science Basis*. Ed. by Valérie Masson-Delmotte et al. Sixth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge: Cambridge University Press.
IPCC Working Group 1 (2013). *Climate Change 2013: The Physical Science Basis*. Ed. by Thomas F. Stocker et al. Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge: Cambridge University Press.
Jebeile, Julie and Anouk Barberousse (2021). Model Spread and Progress in Climate Modelling. *European Journal for Philosophy of Science* 11.66.
Joyce, James M. (1999). *The Foundations of Causal Decision Theory*. Cambridge: Cambridge University Press.

Justus, James (2012). The Elusive Basis of Inferential Robustness. *Philosophy of Science* 79.5: 795–807.

Kass, Robert E. and Adrian E. Rafferty (1995). Bayes Factors. *Journal of the American Statistical Association* 90.430: 773–95.

Katzav, Joel et al. (2021). On the Appropriate and Inappropriate Uses of Probability Distributions in Climate Projections, and Some Alternatives. *Climatic Change* 169.15: 1–20.

Knutti, Reto et al. (2010). Challenges in Combining Projections from Multiple Climate Models. *Journal of Climate* 25.10: 2739–58.

Lloyd, Elisabeth (2015a). Adaptationism and the Logic of Research Questions. *Biological Theory* 10: 343–62.

— (2015b). *Model Robustness* as a Confirmatory Virtue: The Case of Climate Science. *Studies in History and Philosophy of Science Part A* 49: 58–68.

Myrvold, Wayne (1996). Bayesianism and Diverse Evidence: A Reply to Andrew Wayne. *Philosophy of Science* 63.4: 661–65.

— (2017). On the Evidential Import of Unification. *Philosophy of Science* 84.1: 92–114.

O'Loughlin, Ryan (2021). Robustness Reasoning in Climate Model Comparisons. *Studies in History and Philosophy of Science Part A* 85: 34–43.

— (2023). Diagnosing Errors in Climate Model Intercomparisons. *European Journal for Philosophy of Science* 13.20: 1–29.

Parker, Wendy S. (2010). Whose Probabilities? Predicting Climate Change with Ensembles of Models. *Philosophy of Science* 77.5: 985–97.

— (2011). When Climate Models Agree: The Significance of Robust Model Predictions. *Philosophy of Science* 78.4: 579–600.

— (2018). The Significance of Robust Climate Projections. In: *Climate Modeling: Philosophical and Conceptual Issues*. Ed. by Elisabeth A. Lloyd and Eric Winsberg. Cham: Palgrave Macmillan: 273–96.

Parker, Wendy S. and James S. Risbey (2015). False Precision, Surprise and Improved Uncertainty Assessment. *Philosophical Transactions of the Royal Society Part A* 373.3055: 20140453.

Perrin, Jean (1916). *Atoms*. Trans. by Dalziel Llewellyn Hammick. New York: D. Van Norstrand.

Schlosshauer, Maximilian and Gregory Wheeler (2011). Focused Correlation, Confirmation, and the Jigsaw Puzzle of Variable Evidence. *Philosophy of Science* 78.3: 376–92.

Schupbach, Jonah (2018). Robustness Analysis as Explanatory Reasoning. *British Journal for the Philosophy of Science* 69.1: 275–300.

Schwingshackl, Clemens et al. (2019). Regional Climate Model Projections Underestimate Future Warming due to Missing Plant Physiological CO_2 Response. *Environmental Research Letters* 14.11: 1–11.

Sherwood, Steven C. et al. (2020). An Assessment of Earth's Climate Sensitivity Using Multiple Lines of Evidence. *Review of Geophysics* 58.4: e2019RG000678.

Smith, George E. and Raghav Seth (2020). *Brownian Motion and Molecular Reality: A Study in Theory-Mediated Measurement*. Oxford: Oxford University Press.

Stainforth, David A. et al. (2007). Issues in the Interpretation of Climate Model Ensembles to Inform Decisions. *Philosophical Transactions of the Royal Society Series A* 365.1857: 2163–77.

Stegenga, Jacob and Tarun Menon (2017). Robustness and Independent Evidence. *Philosophy of Science* 84.3: 414–35.

Wheeler, Gregory (2012). Explaining the Limits of Olsson's Impossibility Result. *Southern Journal of Philosophy* 50.1: 136–50.

Wheeler, Gregory and Richard Scheines (2013). Coherence and Confirmation through Causation. *Mind* 122.485: 135–70.

Winsberg, Eric (2018). *Philosophy and Climate Science*. Cambridge: Cambridge University Press.

— (2021). What does Robustness Teach us in Climate Science: A Re-Appraisal. *Synthese* 198: 5099–122.